Integrated computer supported editing, approaches and strategies

Botzem, Susanne; Kropac, Ingo H.

Historical Social Research, Vol. 16 — 1991 — No. 4, 106-115

Integrated Computer Supported Editing, Approaches and Strategies

*Susanne Botzem - Ingo H. Kropač**

Abstract: Speaking about »computer tools for editions« most people speak about of the more refined varieties of typesetting systems. On the contrary it can be argued, that the availability of computer based methods for the presentation of huge data objects, could reopen a discussion of what an »edition« actually is. An introduction into such a redefinition is given: an »edition« being defined as a complex of interlinking representations of different degrees of abstraction from the original form of the text. The way in which scanned manuscript, transcribed text and formalized representations of knowledge about the text interrelate, are described.

Today's Situation - Claims for Tomorrow

The traditional technique of editing has become a very sophisticated discipline, which is expensive and takes a long time. The process of producing the edited text cannot be reconstructed because the various traditions of the sources are distributed all over the countries, the special knowledge of the editor often gets lost because there are no means to preserve it. Methodological development sometimes led to a deadlock and the editors make neither head or tail of the increasing number of sources in the Late Middle Ages or the following centuries. Also the main interest of the historical sciences turned away from editing. (1) On the other side commercial oriented scientists copy the sources onto analog (or sometimes digital) storage media and hold any user of their »edition« in tutelage by their selections of parts of the source and a static indictation of them without any criticism. This simple facsimilation of (mostly illustrated) manuscripts cannot be called a critical edition.

* Address all communications to Ingo H. Kropaô, Forschungsinstitut für Historische Grundwissenschaften, Universität Graz, Heinrichstr. 26, A-8010 Graz, Österreich.

So we get confronted with two main problems:

1. Editing is more than classifying, selecting and rendering the sources accessible, editing is based on source criticism. So we have to combine documentary and analytical techniques. The introduction of source-oriented data processing (2) prepared the development of new methods in computer supported editing. So-called *source bases* are defined as machine readable sources, which's content and structure of information are transformed into the formal system of a data base. The various structures of a source demand different processes of transformation as well as different representations on computer system. (3) The most prominent levels of representations are

   - bitmaps to manage sources as a set of images,
   - text bases containing the original text of the source and
   - structured and/or coded information as a formal representation in a factual data base or a data set.

   All possible levels of representation should be realized in a *source bank,* which includes at the same time image bases, textual and factual data bases.

2. Editing does not mean the same to the different disciplines in the humanities. A historical edition differs very much from a linguistic one, even if they are based on the same source, because the disciplines use different rules of editing. Therefore we need a »neutral« representation of the source, in order to apply these different rules in a (semi-)automated way. So each discipline can make use of the same source base to produce an edition of its own.

Producing large-scaled data bases needs a lot of carefulness and responsibility, if the data bases are intended to be used for secondary analysis. Applying the standard of a traditionally printed edition should be the clue of this work.

In order to reach a new view of editing, we have to combine approved techniques of traditional editing together with recent information technology. That means

- interdisciplinary work,
- consideration of the perceptions of the information science and
- transparency of the research process by
  - representing the sources in different levels,
  - automating the processes of transformation between the particular levels and
  - including special knowledge of the editor.

Historical Social Research, Vol. 16 — 1991 — No. 4, 106-115

The Concept of the Integrated Computer Supported Editing

At the *Forschungsinstitut für Historische Grundmssenschaf ten* of Graz university since about five years computer supported technologies have been used to carry out two large editing projects, the *Urkundenbuch der Steiermark und ihrer Regenten* (Chartulary of Styria) (4) and the *Urkundenbuch des Patriarchats Aquileia* (Chartulary of the Patriarcate Aquileia). (5) In order to improve the process of editing, a concept of computer-supported editing has been developed, especially adapted to the requisites of those regional chartularies. The concept has been the basis of an expanded definition of editing: To edit a source means to make available the contents of a source's information as complete as possible in order to enable researchers to use an edition without the need to reference the original. (6) This definition was laid down to seperate this concept against procedures which also are called »computer supported editing« but which make use of the computer to typeset classical printed editions by traditional methods. (7) The concept of the Integrated Computer Supported Editing (8) reaches beyond that to fulfill the claims mentioned above. The goal of the concept is the generation of an information system, which renders accessible

- the source in the different levels of its representation,
- all (automated) procedures of the transformation from one level to the next and
- all decisions and interpretations by the editor to estimate the source.

The management of the different representations of the source as well as the management of declarative and procedural knowlegde of the editor should guarantee transparency for the user, who will be able to reconstruct and revise all decisions in an inter subjective way. It is necessary to implement a system like this as dynamically as possible, because accompaning diplomatic research and computer supported criticism of the source can modify even the data, not only the knowledge base. Whenever we edit or analyse a text critically, we get confronted with the classical trinity of tradition, external and internal features. (9) In order to formalize and copy the evidences of the text as well as the special knowledge of the editor into an information system, it is necessary to classify each information in regard to all decisions of the editorial staff. So we developed a simple, but practicable model of classification:

1. All decisions basing on common sense need no further examination. Information gained by this way is called *immediate information* or *fact.*
2. If an information is gained by a derivation from plain rules, a change of these rules could cause a change of the information. Decisions like

these do not base on common sense but they are specific to a discipline and need a special background knowledge to receive the signals. This kind fo information is called *mediate information* or *derivation* .

3. Some decisions depend highly on the present paradigm of a discipline or the state of research. They only can be reached by special knowledge and the interpretation of a vague set of different information. Everyone using information gained by interpretative acts like these must be able to reconstruct, examine and correct it. We call it *paradigm-based information* or *interpretation.*

Regarding this model a system has been generated, which represents facts and derivations in different data bases and interpretations in a special form of knowledge bases:

1. The *external features* are managed in three representations:
   1.1 Digitized facsimiles of the sources in form of a set of bitmaps show almost all external features as the manuscripts. These facts are the basis for any further work.
   1.2 The transscripted text of the manuscripts - as a paleographical evidence - can be seen as derivation from the bitmaps. It is stored in a »raw text base« which contains coded information about orthography, punctuation, abbreviations, corrections, forms of letter specific for particular words, type of script etc., as it is given in the original source.
   1.3 General paleographical descriptions of the script and of the used letters as well as a documentation of the authentications are stored in the »raw text base«, too. So it is both, a textual also as a factual data base.
2. The *internal features* are managed in the so-called »pre-edited text base«. The textual evidence is similar to traditionally printed editions, information about the formulas of the charters are added in order to ease content analysis.
3. The *tradition* of the charters is added to a factual data base called »repertory«. Similar to an archive's repertory it contains the location of the source, the *regestum* and other information indicating the source.

For each set of sources (i.e. a stock of documents or a municipal chartulary) a *set of images,* a *raw text base,* a *pre-edited text base* and a *repertory* will be generated. The image-data-set as well as the different data bases are linked to the *repertory* in order to access on the one hand the different representations of the source and on the other all information the editor included according to his/her decisions. In this way each sample of sources results in a set of representations of the sources and knowledge

about them. The linkage of the particular *repertories* produces a network of related sources for comprehensive analyses.

Therefore the goals of the concept and the prerequisites for an implementation can be quoted as follows. The procedure has to provide the

- archiving of digitized manuscripts as images,
- improvement of the readability of the (scanned) sources,
- exact transscription of the manuscripts,
- documentation of further external features,
- pre-editing of the transscription,
- integrative management of sources, texts and facts,
- analysis of the internal features,
- representation of declarations and procedures (knowledge base) and
- analyses by a plurality of methods

The realization of this concept should become a new form of documentation for archives and historical research, exemplified by the medieval history of Regensburg.

The Steps of Realization

This year we started the pilot project to the generation of an integrated information system which is intended to be the base for exploring the medieval administration and chancery of Regensburg. The whole project will be carried out in cooperation between the *Institut für Geschichte* of Graz university and the municipal archive of Regensburg.

The realization of this concept mentioned above is determinated by the *Historical Workstation Project.* The application of the KXEICO -system (10) renders possible the management of the different representations and some of the processes of transformation and analyses as mentioned above. To complete the applications software packages called *Standard Format Exchange program (StanFEP)* (11) and *Analytical Semantic Parsing System (ASPS)* (12) will be used in this project. In the following a short description of the particular steps will be introduced by a typical course:

1. The first step is the immediate digitizing of the manuscripts. The scanned material reflect most of the external features. Recent scanner technology reaches high resolutions and graphic libraries allow an effective processing of the images. Within the realm of the high qualitativ image-processing advantages can be listed:
   1. Digitized original sources can be retrieved (Immediate Image Retrieval) by their indication in a factual data base.
   2. Unreadable or rotten parts of charters become readable by interactiv manipulations of gray response curves or the application of fil-

ters.

3. A graphic-editor enables a user to edit any pixel inside the image in order to make clear parts of the script and test his/her »personal« enhancement.
4. In the (hopefully near) future comparisons of patterns, scripts and scribes as well as logical bridges between *parts* of the image and *parts* of the adequate text should improve paleographical research.

The datatype »Imago« of the Κλειω -system enables the linkage between the graphical and the terminological representation of a source. Finally a substantial advantage of this procedure should be stressed: the archiving and application of the digital representation protects the unique transmitted original source.

2. In a second step the external features of the manuscript have to be documented. To generate a *raw text base* is inevitable to describe all those properties of the text which are necessary for the analyses of different disciplines. That means an exact transscription of the manuscripts (paleographical evidence) as well as the documentation of the other external features like the script etc. The paleographical evidence is encoded as follows:
   - The specifical usage of scripts and letters depending on particular words as well as the punctuation of the manuscript is described by a combination of characters, similar to the convention for input and macros of the typesetting system T_EX. (13)
   - The script's size and color, alternations of the ink and the stage of contemporary editing are indicated by start-stop-symbols.

   The description of some external features beyond the textual evidence as general properties of the script and particular selected letters or the size of the document have been formalized and are managed in the factual part of the *raw text base.* Some other information about the manuscript, i.e. the description about the completeness, is stored in a thesaurus system connected to the data base.

3. In the following the encoded raw text has to be »translated« for the distinction of the different disciplines. This step of pre-editing makes use of the internal conversion routines of Κλειω to recode the text into a form historians are used to. At the present the use of capital letters and punctuation has to be altered manually. Also information about the formulas of the charters are included by the editor using respective tags. Therefore the pre-edited text base consists of text, splitted into various elements (formulas), in order to analyse the charter's formal structure. In this way the text will be prepared in the *pre-edited text base* for further printed edition and can be used for the analysis of internal features by different methods.

4. The next step is the generation of the *repertory*. For the production of the *regestum* the text is explored, information about date and place of issue as well as prosopographical information will be gained from the pre-edited text using ASPS and transformed into Κλειω's input format. (14) The form of the source's tradition is stored now in a thesaurus as a starting point for a semantic network over the whole information system.

The very last step of the concept provides for the computer supported *discrimen veri ac falsi.* The result is the final estimation of the source, gained by comprehensive analyses in quantitive and qualitative way.

Summary

In this way the project provides or will provide

- a model of an information system for historically oriented disciplines, exemplified on the medieval history of Regensburg,
- an archive of a large amount of digitized original medieval sources,
- the improvement of the readability of these handwritten manuscripts by applying the benefits of image processing,
- integrated managing of images, continuous text and structured facts
- logical bridges between parts of the texts and parts of the images and
- the development of basic (computer-supported) methods and techniques in the history of chanceries in the Late Middle Ages.

Notes

(1) See Bruno Meyer: Zur Edition historischer Texte. In: Schweizerische Zeitschrift für Geschichte 1, 1951, pp. 177 - 220; Horst Fuhrmann: Über Ziel und Aussehen von Texteditionen. In: Mittelalterlicher Textüberlieferungen un ihre kritische Aufarbeitung, ed. by Horst Fuhrmann. München 1982, pp. 12-27; Carlrichard Brühl: Derzeitige Lage und zukünftige Aufgaben der Diplomatik. In: Landesherrliche Kanzleien im Spätmittelalter, ed. by Gabriel Silagi. München 1984 (Münchner Beiträge zur Mediävistik und Renaissance-Forschung 3),pp 37^17.

(2) See Manfred Thaller: Vor über legungen für einen internationalen Workshop über die Schaffung, Verbindung und Nutzung großer interdisziplinärer Quellenbanken in den historischen Wissenschaften. In: Datenbanken und Datenverwaltungssysteme als Werkzeuge historischer Forschung, hrsg.v. Manfred Thaller. St. Katharinen 1986 (Historisch-Sozial wissen schaftliche Studien 20), pp. 9-30; Manfred

Thaller: The Daily Life in the Middle Ages, Editions of Sources and Data Processing. In: Medium Aevum Quotidianum Newsletter 10, 1987, pp.6-28.

(3) See Ingo H. Kropaé: Gain et perte d'information. Problêmes fondamentaux pose's par l'édition informatisée de donnée̊s historiques. In: Standardisation et échange des bases de données historiques. Actes de la troisième Table Ronde internationale tenue au L.I.S.H. (C.N.R.S.), ed. by Jean-Philippe Genet. Paris 1988, pp. 49-57.

(4) See the recent report on the project: Friedrich Hausmann: Urkundenbuch der Steiermark und ihrer Regenten, Band I - III und V ff. In: XXII. Bericht der Historischen Landeskommission für Steiermark, 16. Geschäftsperiode (1982 - 1986), ed. by Othmar Pickl. Graz 1988, S. 79-90.

(5) The first volume of this project is already published: Reinhard Härtel: Die ältesten Urkunden das Klosters Moggio (bis 1250). Wien 1986 (Publikationen des Historischen Instituts beim Österreichischen Kulturinstitut in Rom, 2. Abt., Reihe 6 = Vorarbeiten zu einem Urkundenbuch des Patriarchats Aquileia 1). For a short survey of the project see Reinhard Härtel: Il progetto di ricerca e di edizione »Urkundenbuch des Patriarchats Aquileia« (Codice diplomatico del patriarcato d'Aquileia). In: Memorie storiche forgiuliesi 61, 1984, pp. 177-186.

(6) See Ingo H. Kropad: Informationssysteme in der Geschichtswissenschaft. Konzeption und Anwendung am Beispiel der Prosopographischen Datenbank zur Geschichte der südöstlichen Reichsgebiete bis 1250 (PDB). Graz 1988 (dzt. unveröff. Habilschrift), pp. 53-59; Reinhard Härtel - Ingo H. Kropaà: Edition und Auswertung mittelalterlicher Urkunden. Probleme bei Standardisierung und Transfer fortlaufender Texte. In: Data Networks for the Historical Disciplines? Problems and Feasibilities in Standardization and Exchange of Machine Readable Data, ed. by Friedrich Hausmann et al. Graz 1987, pp. 100-112; Ingo H. Kropad: Homo ex Machina: Prosopography and Chartularies. In: Data Base Oriented Source Editions. Papers from two sessions at the 23rd International Congress on Medieval Studies, Kalamazoo, Michigan, 5-8 May, 1988, ed. by Manfred Thaller, pp. 37-45; Ingo H. Kropad: Quellenbanken als Editionsmedien und ihre Rolle in fach spezifisch en Informationssystemen. In: Historische Edition und Computer. Möglichkeiten und Probleme interdisziplinärer Textverarbeitung und Textbearbeitung, ed. by Anton Schwöb et al. Graz 1989, pp. 243-262.

(7) See Wilhelm Ott - Hans Walter Gabler - Paul Sappler: EDV-Fibel für Editoren. Stuttgart 1982, pp. 25-34; Hans Walter Gabler - Wolf gang Kreitmair: Der Computer als Arbeitshilfe für das wissenschaftliche

Edieren. In: Computerfibel für die Geisteswissenmschatten. Einsatzmöglichkeiten der Personal Computers und Beispiele aus der Praxis, ed. by Bernd Gregor und Manfred Krifka. München 1986, pp. 203-211; Susan Hockey: A Guide to Computer Applications in the Humanities. London 1980, pp. 144-167; Robert L. Oakman: Computer Methods for Literary Reserach. Columbia 1980, pp. 113-138; Wilhelm Ott: Computer Applications in Textual Critcism. In: The Computer and Literary Studies, ed. by A.J. Aitken, R.W. Bailey, N. Hamilton Smith. Edinburgh 1973, pp. 199-206.

(8) See Susanne Botzem und Ingo H. Kropac: Integrierte Maschinelle Edition am historischen Arbeitsplatzrechner: Repräsentation und Dokumentation von Quellen und Wissen am Beispiel des Regensburger Kanzlei-und Verwaltungswesens im Spätmittelalter. Graz, Institut für Geschichte 1989 (Integrierte Maschinelle Edition - Bericht 1, un ver - öff. Arbeitspapier).

(9) Explicitly stressed by Friedrich Leist: Urkundenlehre. Katechismus der Diplomatik, Paläographie, Chronologie und Sphragistik. Leipzig 1893.

(10) Manfred Thaller: **Κλειω** 3.1.1. Ein Datenbanksystem. St. Katharinen 1989 (Halbgraue Reihe zur Historischen Fachinformatik, Serie B: Softwarebeschreibungen, Bd.l); Ingo H. Kropac und Ursula Leiter-Köhrer: Kteuo. Der Datenbankeditor. St. Katharinen 1989 (Halbgraue Reihe zur Historischen Fachinformatik, Serie B: Softwarebeschreibungen, Bd. 5); Peter Becker: **Κλειω** . Ein Tutorial. St. Katharinen 1989 (Halbgraue Reihe zur Historischen Fachinformatik, Serie A: Historische Quellenkunden, Bd. 1). In the following we shall make use of the terminology of the **Κλειω** -system.

(11) Manfred Thaller: A draft proposal for a format exchange program. In: Standardisation et échange des bases de donnees historiques. Actes de la troisième Table Ronde internationale tenue au L.I.S.H. (C.N.R.S.), ed. by Jean-Philippe Genet. Paris 1988, pp. 329-375.

(12) The project to develop this system - sponsored by the Austrian *Ministerium für Wissenschaft und Forschung* - is carried out at the *Forschungsinstitut für Historische Grundwissenschaften* at Graz university. See Ingo H. Kropaé - Ursula Leiter-Köhrer: Analytical Semantic Parsing System: ein Programm zur automatisierten Indizierung und Inhaltserschließung. Graz 1989 (ASPS Bericht 1, unveröff. Arbeitspapier)

(13) See Donald E. Knuth: The TEXbook. Reading etc. 1984.

(14) The data model is closely related to the Prosopographical Data Base at Graz university. See Ingo H. Kropac und Peter Becker: Die Prosopographische Datenbank zur Geschichte der südöstlichen Reichsgebiete bis 1250 (PDB). Konzepte und Kurzdokumentation. In: Medium

Aevum Quotidianum Newsletter 10,1987, pp. 30-47; Ingo H. Kropac: Who's who in the Southeast of Germany. The Design of the Prosopographical Data Base at Graz University. In: History and Computing II, ed. by Peter Denley, Stefan Fogelvik, Charles Harvey. Manchester-New York 1989, pp. 273-279.